Comparison of Cancer Cell Line Genomic Profiling Data Reported by Sanger and CCLE

Zongfu Cao[1], Xiaoqiao Liu[1§], Yue Wang[1]

1. Department of Quantitative Biology, BeiGene (Beijing) Co., Ltd, Beijing 102206, China.

§ Corresponding author: xiaoqiao.liu@beigene.com(Xiaoqiao Liu)

Abstract

Background

Cancer cell lines have been widely used as the pre-clinical model system to evaluate the impact of drug candidates on the cell level activities. The availability of systematic molecular profiling data for cancer cell lines also provides a great opportunity to evaluate the molecular basis of drug activity. As of today, Sanger Cancer Cell Lines Project (Sanger) and Broad-Novartis Cancer Cell Line Encyclopedia project (CCLE) are two initiatives that provide the latest and largest published data sets on the baseline profiling for cancer cell lines. Though some comments emerged for these two related works, no detailed comparison has been provided for these two datasets. Here, in order to examine the degree of concordance and maybe give some caveats for application, we compared the most up-to-date data on mRNA expression, DNA copy number and DNA mutation reported by Sanger and CCLE.

Methods

For expression and copy number data, we base our comparison on gene-centric data reported from each project. For mutation data, we use the “Preferred dataset” from CCLE (which concentrates on non-neutral, CDS mutations) and all mutation data from Sanger. Altogether, 12,012 genes at expression level, 362 genes at copy number level and 60 genes at mutation level are overlapped for more than 440 cell lines from the two projects. Pearson correlation coefficient and Jaccard index are reported here as measure for consistency. Detailed comparison for three types of data sets at cell line and gene level is also provided respectively.

Results

1. For mRNA expression data, the correlation is considerably strong, with overall Pearson correlation coefficient 0.902. Besides, cell lines from two data sets possess similar clustering pattern.
2. The correlation for copy number data is relatively high, with overall Pearson correlation coefficient 0.626.
3. For mutation, the consistency for two data sets is not as good as copy number or expression data, mostly likely due to different mutation calling strategy. Totally, about 35.2% mutations are reported by both Sanger and CCLE with the criterion of identical cell line name, chromosome position and alteration of genomic sequence.

Discussions

The difference between the profiling data from CCLE and Sanger may be caused by several reasons. Besides cell culture conditions, discrepancy for profiling methods may come from the following aspects.

1. For mRNA expression data, different microarray platforms with different probe sets for the same gene may contribute to the difference for mRNA expression data.
2. For copy number data, Sanger uses PICNIC-based and CCLE uses GenePattern-based strategy, which would give relatively different estimation for copy number. Besides, different normal samples used by each project also need to be considered.
3. For mutation data, the inconsistency may come from two aspects: (1) different sequencing platforms, capillary sequencing used by Sanger and targeted massively parallel sequencing used by CCLE. (2) Mutation filtering strategies: considerably different protocols are used by two projects and may be the main factor to explain the inconsistency of mutation data.